

서처와 정보검색

Searcher and Information Searching



장재경
한국소비자보호원
Chang, jae-gyong
Consumer Protection Board

세계의 정보가 손끝에 있다는 말은 과언이 아니다. 그러나 세계 도처에는 수십억건에 육박하는 정보들이 홍수처럼 불어나고 있고, 수천여 데이터베이스의 다양한 정보내용과 구성포맷, 접근방식 및 정보이용료 등을 파악해야 하는 한편 빠르게 바뀌고 있는 데이터베이스 산업 동향을 제대로 파악하기는 용이한 일이 아니다.

급변하는 정보화사회에서 필요한 정보를 빠른시간에 획득하여 효과적으로 활용하는 것이 중요한 성공전략으로 여겨지고 있음에 따라, 필연적으로 서처에 대한 사회적 요구가 증대되

고 있다.

외국에서는 'Searcher'라는 말 외에 'Reference librarian'이나 'Information broker'라는 말이 사용되는데 각각의 말에서 주는 뉘앙스와 같이 다소 그 역할에는 차이가 날 수 있지만, 기본업무는 주로 정보를 찾는 일을 한다. 혹은 경우에 따라서 서처를 정보스파이, 산업스파이로 곡해하기도 하는데, 공개된 정보를 원칙적인 대상으로 하고 있다는 점에서 차이가 있다고 하겠다. 미국이나 유럽에서는 실제로 많은 서처들이 10년 이상의 도서관 사서 경력을 갖고 독자적인 주제분야를 갖추

어 독립해서, 프리랜서 형식으로 일을 하는 정보베테랑이라 할 수 있다.

미국의 AIIP(Association of Independent Information Professionals), 유럽의 EIRE-NET(European Information Researchers Network)과 같은 공식적인 서처모임이 있다. 이들 중 특히 필요한 정보를 전문적으로 찾아내는 정보브로커는 고급 전문직으로 대우받고 있다. 한편 일본에서는 국가인증 자격제도를 운영하여 1급, 2급 서처자격을 인증하고 있다. 많은 일본기업체에서는 이들을 주로 全社的 고급정보의 수집

및 기획에 주력한 정예팀으로 구성하여 회사의 전략(기획)정보를 전담하도록 한다.

서처는 문제해결을 위해 필요한 정보를 신속 정확하게 찾아내는 일에서 부터 찾은 정보로부터 새로운 부가가치를 창출해내는 일까지 할 수 있다. 가장 적합한 정보소스를 파악하여 역세스하고 특정분야의 전문정보를 해석하고 분석할 수 있는 정보수집의 첨병으로서의 역할을 한다. 즉 어떠한 데이터베이스에 접근하여야 하는지, 원하는 정보를 입수하기 위한 정보검색 전략을 세우고, 실제로 신뢰할 만한 정보를 입수하는 일, 그 정보를 분석하여 경우에 따라서는 예측정보나 전략정보 등을 창출해내는 일이라 하겠다. 국내에서도 국가의 각종 데이터베이스 산업육성지원에 따라 급속히 데이터베이스산업이 발전하고 있으며 정보화사회를 향한 정보마인드의 확산되고 있어서 향후 조속히 서처의 활발한 활동이 요구되어질 것으로 보인다.

(1) 정보의 Know Where과 Know How to Get

필요한 정보가 어디에 있는지를 아는 것은 몹시 중요한 문제이다. 그러나 세계 곳곳에서 끊임없이 새로운 유형의 정보가 생산되거나 혹은 생산이 중단되

고 있으며, 수천개의 상용 데이터베이스에서부터 인터넷 구석 구석에 들어 있는 어마어마한 정보까지 일목요연하게 정리되어 있지도 않다. 그리고 각 정보들도 형형색색 독자적인 특성을 갖고 있으며 대체로 검색방법과 주제가공 형식이 통일되어 있지도 않다. 따라서 서처는 각종 정보의 소재를 총괄적으로 파악하고 있어야 한다. 한편 그 정보에 어떻게 접근하여 원하는 정보를 입수할 수 있을 것인지를 또한 익혀두어야 한다. 이는 통신망을 통할 것인지, CD-ROM과 같은 매체를 통할 것인지 인쇄 매체를 구해야 할 것인지 등에 대한 다각적인 지식을 갖추고 있어야 한다 이는 결국 입수비용과 연결되며 동일한 정보도 각 매체 마다 정보의 내용(예를 들면 서지사항, 초록, 전문 등)이 다를 수 있으므로 이를 사전에 파악하여야 한다.

● 이용하여야 할 정보망은?

Online誌에서 설문조사한 결과 서처들이 가장 많이 이용하는 정보망은 바로 Dialog였다. Dialog가 서처들에게 가장 많이 이용되고 있는 이유는 원하는 검색을 다양한 방법으로 정확하게 검색해 볼 수 있다는 점이다. 익숙해 지기에는 다소 시간이 걸리고 불편하지만 일단 익숙하

고 나면 오히려 원하는 정보를 쉽게 찾을 수 있다.

Dialog는 1974년 NASA에서 추진한 프로젝트의 일환으로 개발이 되었던 터라 비교적 과학 기술, 의학분야의 정보가 강하며 수년전 부터 비즈니스쪽 정보가 강화되었고 1993년 유럽의 Datastar와 합병이 되어 보다 방대한 정보를 서비스하게 되었다. 반면 법률 분야의 정보가 취약한 편이다. 법률 분야의 정보 접근을 위해서는 Lexis/Nexis를 이용하는 것이 좋다. 미국의 CompuServe는 금융, 재무, 신용, 현물정보가 강한편이다. 작년 유럽EC에서는 프랑스의 Questel를 중심으로 Orbit와 BRS가 통합되었고 이는 특히 유럽의 과학기술 분야가 강하다. 일본의 정보는 NiftyServe나 PC-VAN을 통하여 이용할 수 있으며, 일본경제신문사의 니케이텔레콤을 이용할 수 있다. 최근 세계각국의 정보망이 모이고 있는 인터넷은 방대한 도서정보인 미국 의회도서관 정보나 OCLC의 First Searching 외에 각종 공공정보 및 각종분야의 전문정보, 취미·오락정보가 있으며 최근 CompuServe, America Online, Prodigie 및 프랑스의 생활정보망인 Minitel과 같은 상용망이 인터넷에 연결을 추진하고 있다.

지난해에 출현한 Commerce-



Net은 인터넷상에서 전자상거래를 지원하기 위한 기반구축 등을 꾀하고 있어 인터넷은 앞으로 보다 거대한 세계망으로 계속 확대될 것이다.

● 신뢰할 수 있는 정보를 찾아야

서치는 우선 믿을 만한 정보를 찾아야 하므로 원정보의 소스에 유념하여야 한다. 가장 공신력이 있는 정보로는 제도적 측면이 강한 법률정보, 각국의 정책정보, 특허정보가 있고 반면 뉴스기사 또는 각종 분야의 게시물과 같은 정보는 상대적으로 신뢰성이 떨어진다고 하겠다. 그리고 과학, 기술의 빠른 발전으로 인해 정보의 노화가 가속되고 있으므로 서치는 최신정보를 누락하는 일이 없도록 하여야 한다. 미국에서는 잘못된 정보를 서비스함으로써 종종 분쟁이 야기되는데 이를 위한 손해 보상보험 제도 설치를 추진한다고 한다.

● 원하는 정보를 얼마나 찾을 수 있을까?

선진국의 경우 과학, 비즈니스 정보 등의 대부분 정보는 80% 정도 데이터베이스화 되어 있고 생활정보는 아직까지 30% 내외라고 한다. 미국은 정보공개법(The act of Informtion)이

시행되고 있어 의회 회의록에서부터 정부조달정보까지 갖가지 정보를 얻을 수 있고, 심지어 TV드라마 대본까지도 구해볼 수 있다. 반면 유럽 EC제국은 통신에 관한 제약을 법으로 설정하고 있어 자국 정보의 유출이 까다로운 편이다. 그밖에 정보후진국의 정보나 언어권이 다른 지역의 정보는 비교적 데이터베이스화율이 떨어지고 있지만 실제로 정보선진국에서 이를 대상으로 수집한 정보를 중심으로 부족하나마 이를 커버해 주고 있다. 예를 들면 미국 의회도서관에서는 우리나라에서 발행된 도서까지 오래전부터 적극적으로 수집하고 있어 국내 어느 도서관보다 우리나라의 관련자료를 충실하게 소장하고 있다고 한다. 최근 정보의 출판에도 컴퓨터가 활용되어 데이터베이스화 비율이 점점 향상되는 한편 많은 정보가 쏠문으로 제공되고 있다. 그리고 세계 구석구석에 걸쳐 인터넷에 기하급수적으로 정보가 연결되고 있기 때문에 서치가 이용가능한 정보는 계속 빠르게 증가될 전망이다.

● 데이터베이스검색을 보조하는 디렉토리나 서칭툴은?

데이터베이스마다 정보의 수록 범위 및 주제가공 양식 등에 차이가 있기 마련이다. 그리고

세계에 유통되고 있는 약 8,000개의 데이터베이스 가운데 계속 새로운 정보가 생산되는 반면 생산중지 되는 정보가 있기 때문에 이를 한꺼번에 파악하기는 용이치 않다. 따라서 이를 위해 데이터베이스에 관한 디렉토리나 서칭툴을 이용하는 것이 좋다. Caudra/Elsevier社에서 발행한 「Online directory」나, 일본의 동양공업신문사에서 발행한 「온라인디렉토리」가 있다. 한편 방대한 인터넷 정보에 대한 접근을 위해서는 「WWW Yellow Page」, 「Netscape Galleria」, 「Virtual Library」, 「Virtual Town」, 「Yahoo」, 「Global Network Navigation」 등의 서칭툴이 있다. Dialog에서 제공되는 각 데이터베이스의 상세한 정보내역 및 필드, 접두어(Prefix), 접미어(Suffix) 등을 기술한 「Blue-sheet」와 Dialog에서 제공하고 있는 전레코드의 색인어를 연결한 「Dialindex」가 있어 Dialog의 수억개의 레코드 중 찾고자 하는 특정정보가 어떤 데이터베이스에 가장 많이 수록되어 있는지를 랭킹하여 주는 몹시 편리한 기능을 제공하여 준다.

● 정보이용요금

최근 정보통신망의 발달로 정보망 접속료는 대체로 낮아지고



있으나, 정보이용요금은 크게 떨어지지 않는다고 있으며 내용이 계속 보강되고 있는 특허정보나,全文서비스가 가능해진 일부 데이터베이스의 정보이용 요금이 오히려 조금씩 늘어가고 있어 정보화사회의로의 진행의 한 면모를 볼 수 있다. 인터넷이 세계 유일의 거대한 망으로 등장할 수 있었던 결정적인 요인은 바로 저렴한 비용 때문이었다. 인터넷에 수록된 정보는 정보공유를 위한 정보마인드가 일찌기 발달되었던 미국에서 자발적인 정보제공자의 의욕에 따라 제공되거나 혹은 정보공개원칙으로 각종 도서관 소장정보나 국가의 행정, 법률, 정책 등 관련 공공 정보 등으로 값싼 정보의 공유

채널로서 성장해 왔다. 그러나 작년부터 인터넷이 상용화 추세로 가고 있기 때문에 미국의 CompuServe, America Online, Prodigy 등 정보망 접속이 가능하게 되었다. 이는 결과적으로 정보망의 접속료의 부담을 계속 떨어뜨리도록 하는 데 공헌하였다. 한편 Dialog는 미국의 대형 학술정보망인 OCLC의 「EPIC」과 「First Searching」과의 경쟁의 일환으로 학생이용자를 대상으로 Dialog의 데이터베이스 일부를 저렴하게 제공하고 있는 「Knowledge Index」를 개설하였다. 그러나 엄밀한 의미에서 「Knowledge Index」의 정보는 연구를 지원하기 위한 서지정보 또는 초록

정보 위주로 Dialog와 차별화된 상품이라고 할 수 있다.

(2) 온라인정보검색과 CD-ROM 검색

수년전 부터 정보서비스 산업에서 CD-ROM시장이 급격히 발전을 해오고 있으나 아직도 온라인시장이 매출의 5분지 3 이상을 주도하고 있다. 통신망을 이용한 온라인검색의 최대장점은 신속성에 있으며 제공정보의 종류가 CD-ROM에 비해 다양하므로 정보검색의 주요대상이라고 할 수 있다. 즉 온라인정보는 최신 Realtime 뉴스에서부터 Stock 데이터베이스까지 다양하게 이용할 수 있는데다



양방향통신이 가능한 점에서 적극적 의미의 정보수집 채널이기도 하다. 하지만 CD-ROM은 일시적 구입비만 내면 요금 부담이 없이 계속 이용이 가능하기 때문에 특정분야의 정보이용빈도가 높은 경우에는 오히려 편리하고 경제적이라고 할 수 있겠다. 그리고 온라인검색에서 지원이 어려웠던 화상정보를 CD-ROM에서는 지원할 수 있기 때문에 최근 비즈니스와 의학분야에서 각 분야의 저널을 집대성한 서지초록 정보와 함께 원문을 그대로 제공하는 「BPO」, 「GPO」, 「ADONIS」 등의 이용도 편리하다.

그리고 CD-ROM정보를 많이 이용하는 분야로 특히 도면 등의 화상정보가 중요한 의미를 갖게 되는 특허정보가 있으며, CD-ROM의 막대한 정보저장능력에 따라 백과사전이나 학위논문, 도서관서지정보, 의학정보 등도 많이 이용되고 있다.

최근 인터넷의 WWW를 통해 멀티미디어 정보서비스가 가능해진 했지만 화상데이터의 전송에는 통신망속도에 제한이 있는 편이다.

● 원하는 정보를 찾으려면

정보망에 접근하기 전에 충분한 준비를 하지 않으면 쓸데없이 큰 비용을 낭비하기 쉽다. 각 정

보망에 따른 정보의 액세스방법이 다르거나 각 데이터베이스마다 주제가공서비스 체제가 다르기 때문에 비싼 정보이용료를 감안해야 하는 온라인 검색에 있어서 충분한 사전 준비가 없이는 당황하지 않을 수 없다. 그러한 점을 보완하기 위해 검색방식이 다른 Dialog, Newsnet, Questel 등에 수록된 모든 정보를 통합하여 인공지능을 이용하여 검색지원을 해주는 Easynet 정보망이 있다. 이는 편리하긴 하지만 검색비용이 추가적으로 소요되어 비싼게 흠이다. 서치가 원하는 정보를 찾으려면 정보망에서 제공되는 데이터베이스에 대한 정보내역과 검색원리를 상세하게 파악하고 있어서 실제 검색시 각종 검색소프트웨어가 제공하는 기계적인 지원한계를 간파해낼 수 있어야 한다. 예를 들면 「Dialindex」를 통하여 'Electric Car'를 검색하였을 때 만약 「ABI Inform」과 「NTIS」라는 데이터베이스에서 기계적으로는 동일하게 히팅되었다고 해도 실제정보는 데이터베이스에 따라 전혀 별개의 내용이 들어 있을 수 있기 때문에 이를 파악하여 원하는 데이터베이스로 액세스하는데 착오가 없도록 하여야 한다. 그리고 검색시 자연언어로서 색인된 데이터베이스 인지 아닌지를 미리 알아보아 해당 데이터베이스에 사

전에 부여된 수준의 검색어에 맞추어서 검색하여야 한다. 그래서 서치가 실제 데이터베이스에 1시간 동안 검색을 하는데는 10시간의 사전준비가 따라야 한다고 한다.

● 검색의 ABC

정보검색 방법은 크게 세가지로 구분된다.

1. 메뉴방식

이용자에게 먼저 메뉴를 제시하여 이용자가 필요한 메뉴까지 접근하여 필요한 정보를 찾아내는 방식인데, 우선 쉽기는 하지만 데이터량이 일정분량을 초과하는 경우에는 오히려 복잡하고 메뉴를 따라 접근해서 필요한 정보를 찾는 수동적인 검색방법이다. 새로운 뉴스가 계속 로딩되며 일정기간의 정보만을 보관하는 Realtime 데이터베이스나 각종 생활안내 정보 등에서 유용하다.

2. 명령어검색

정보의 양이 많은 Stock 데이터베이스에서 원하는 데이터를 쉽게 검색할 수 있기 때문에 서처에게 선호되는 방식으로 일반 이용자인 경우에는 어느 정도 숙달되어야 쉽게 이용할 수 있다.

명령어 검색은 이용하기 전에

이용하고자한 데이터베이스가 어떻게 주제가공 되었는 지를 미리 살펴보는 것이 보다 효율적인 검색을 할 수 있다. Dialog, STN, Questel 등 많은 정보망에서 명령어검색을 지원하고 있다.

3. 하이퍼텍스트 검색

새로운 GUI환경에서의 정보 검색을 위해 최근 많이 이용하고 있는 Mosaic, Netscape와 같은 Web Browser에서 HTML (HyperText Markup Language)로 지원된 하이퍼텍스트 검색은, 하이라이트된 단어를 마우스로 클릭하기만 하면 연관된 정보와 연결되는 검색으로 무척 손쉬운 검색방법이다. 하지만 정보항해(Navigation)시 방향 감각을 잃어버리기 쉬우며(Disorientation), 원하는 정보를 진입하기 위해서 필요없는 정보노드를 거쳐야 하는 점(Cognitive Overhead)과 아직 Boolean 검색을 병행할 수 없다는 단점을 지니고 있으나 멀티미디어 환경에서의 효율적인 정보검색 도구로서 향후 많은 연구가 기대된다.

● 검색효율을 높이기 위해서

검색시 관련된 정보를 잘 검색해내는 정도를 재현율(Recall)이라 하고, 정확하게

필요한 정보를 찾아내는 정도를 정확율(Precision)이라 하는데, 재현율과 정확율은 서처의 검색 결과에 있어서 무척 중요한 평가척도가 되고 있다.

본고에서 우선 재현율과 정확율을 합한 총체적 의미의 검색 효율을 높이는 방안을 간략히 살펴보면,

1. 불리언(Boolean) 연산자

검색하고자 하는 개념을 AND와 OR의 조합인 불리언 논리식으로 표현하는 것으로 간단히 예를 들면 '미국의 전기자동차' 라는 개념은 '미국 AND 전기자동차' 로 검색식을 세울 수 있다. 하지만 특히 자연언어로 자동색인된 데이터베이스를 검색시에 유의할 점은 동의어를 AND 연산자로 충분히 확장시켜 주어야 한다는 점이다.

2. 필드검색

필드를 충분히 활용하는 것은 검색에 중요한 영향을 미칠 수 있는데 특히 중요하게 파악해야 할 필드로는 해당 데이터베이스의 주제를 가공한 필드이다. 즉 해당 데이터베이스 특유의 주제 분류코드 필드, Descriptor, Subject Heading 등 주제가공 필드를 충분히 활용하도록 해야 한다. 한편 회사명, 저널명, 저자명 등의 검색시 노이즈 발생을 줄이기 위해 필드검색으로

확인하는 것도 중요하다.

3. 자연어검색 VS. 통제어검색

최근에 개발된 많은 데이터베이스는 자연어색인에 의해서 서비스되고 있으며, 전통적인 대형 데이터베이스는 시소러스(Thesaurus)를 이용한 색인어가 들어있는 경우가 많다. 검색 효율을 높이기 위해서 자연어검색인 경우 동의어를 많이 활용하여 검색하는 것이 필요하고 통제어검색인 경우 색인필드를 충분히 활용하는 것이 효율적이다. 필자의 경험으로 비추면 통제된 색인어가 있다면 100% 활용하는 것이 검색의 비용과 노이즈를 줄이는 현명한 방법이었다.

4. 句(Phrase)검색과 근접

연산자(Primitive Operator)

데이터베이스 구축이나 검색에 있어 의외로 까다로운 것이 띄어쓰기 문제이다. 띄어 쓰기를 하는 복합어가 많은데 이때 구검색을 하면 편리하다. 또한 새로운 개념의 검색이나 연관성이 있는 어떤 개념을 검색을 하고자 할 때 유용하게 활용할 수 있는 것이 근접연산자이다. 우리나라 데이터베이스구축에는 많이 활용되지 않아 생소할 지는 모르지만 몹시 유용한 기능이다.

5. 용어처리

유사어를 효과적으로 찾기 위

하여 용어절단(Truncation)을 활용한다. 특히 영어의 경우 복수처리 및 어미변화가 있으므로 적절한 용어처리는 필수이다.

● 시소러스를 활용하려면

시소러스는 용어(색인어)간의 관계를 설정한 주제어사전이다. 시소러스의 개발 및 최신성 유지에는 많은 비용이 들기 때문에 역사가 깊은 대형 데이터베이스에 주로 시소러스가 지원되고 있다. 주로 이용되고 있는 시소러스는 Medline의 MeSH, Eric, Root, Inspec, 일본과학기술정보의 JICST, 일본경제신문의 日經시소러스가 있다. 데이터베이스를 검색할 때 관련데이터베이스에 관한 시소러스를 책자형으로 구비하고 있으면 편리하나, 필요한 부분을 온라인으로 열람해 볼 수도 있다. 시소러스가 내장되어 있는 경우는 시소러스에서 지정한 디스크립터를 통해 주제검색하는 것이 가장 효율적이다.

(3) 정보의 부가가치 창출

정보화시대로의 진행속에서 단순히 정보를 검색하는 일 외에 기존의 정보를 이용하여 각종 부가가치가 높은 새로운 정보를 창출하는 것이 중요한 테마가 될 수 밖에 없다. 바야흐로

세계는 지구촌이라는 개념으로 축소되고 있으며 따라서 기업의 경쟁력을 확보하기 위해 정보를 효율적으로 활용하여 새로운 이익을 창출코자 하는 부단한 노력이 정보화시대의 성공전략이 아닐 수 없다.

따라서 어떠한 문제해결을 위하여 필요한 정보를 적극적으로 수집하여 계량적으로 분석함으로써 새로운 돌파구를 마련할 수 있을 것이다.

● 예측정보, 의사결정정보 서비스

정보화사회 속에서 기업은 보다 다양한 소비자의 기호 및 요구를 충족시키기 위하여 다품종 소량 생산체제를 지향하게 될 것이며, 관련하여 많은 의사결정이 필요하게 된다. 따라서 검색한 정보를 계량경영학적 방법론으로 분석하여 수요예측정보, 투자예측정보, 의사결정지원 정보를 생산해 낼 수 있다. 실제로 미국의 왓슨경영연구소나 각종 경영컨설팅사는 그러한 정보를 토대로 새로운 부가가치를 창출하고 있다.

● 전략정보, 경쟁정보 서비스

한편 기업이 당면한 현실은 무수한 경쟁 속에서 전략우위를 차지해야 하는 점이다. 그러한

환경에서 경쟁사에 대한 추적 정보의 운영 또는 새로운 이익을 창출하기 위한 투자전략 정보, 상품개발 전략정보, 경영전략정보 등이 필요하다. 이를 위해 세계적 규모의 정보망을 통해 필요한 정보를 적시에 분석하여 전략적 활용을 목적으로 하는 새로운 부가가치를 창출할 수 있다.

따라서 서처는 국내 정보는 물론 해외 구석구석의 정보까지도 필요시 신속하게 입수할 수 있도록 하여야 하며 아울러 이를 분석하고 해석할 수 있는 안목을 갖추어서 향후 정보화사회의 첨병으로써 커다란 몫을 할 수 있을 것이다. 그러나 국내에서 아직 서처를 인증하는 제도는 마련되지 않고 있다. 

